Act Like a Radiologist: Towards Reliable Multi-view Correspondence Reasoning for Mammogram Mass Detection

Yuhang Liu*, Fandong Zhang*, Chaoqi Chen, Siwen Wang, Yizhou Wang and Yizhou Yu, *Fellow, IEEE*

Abstract—Mammogram mass detection is crucial for diagnosing and preventing the breast cancers in clinical practice. The complementary effect of multi-view mammogram images provides valuable information about the breast anatomical prior structure and is of great significance in digital mammography interpretation. However, unlike radiologists who can utilize the natural reasoning ability to identify masses based on multiple mammographic views, how to endow the existing object detection models with the capability of multi-view reasoning is vital for decision-making in clinical diagnosis but remains the boundary to explore. In this paper, we propose an Anatomy-aware Graph convolutional Network (AGN), which is tailored for mammogram mass detection and endows existing detection methods with multi-view reasoning ability. The proposed AGN consists of three steps. Firstly, we introduce a Bipartite Graph convolutional Network (BGN) to model the intrinsic geometric and semantic relations of ipsilateral views. Secondly, considering that the visual asymmetry of bilateral views is widely adopted in clinical practice to assist the diagnosis of breast lesions, we propose an Inception Graph convolutional Network (IGN) to model the structural similarities of bilateral views. Finally, based on the constructed graphs, the multi-view information is propagated through nodes methodically, which equips the features learned from the examined view with multi-view reasoning ability. Experiments on two standard benchmarks reveal that AGN significantly exceeds the state-of-the-art performance. Visualization results show that AGN provides interpretable visual cues for clinical diagnosis.

Index Terms—Detection, graph convolutional network, reasoning, multi-view, mammogram.

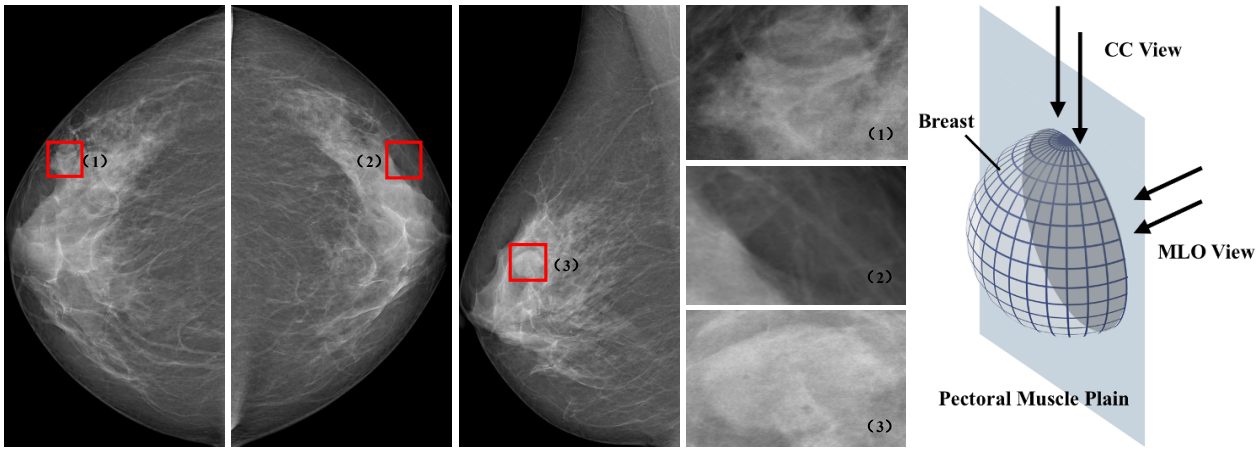
1 Introduction

Breast cancer, which has the highest incidence and mortality rates among women [1], is one of the leading cause of cancer deaths worldwide. Screening mammography has demonstrated strong efficacy in reducing the breast cancer mortality especially at the early stage [2]. The detection of masses based on mammograms is a key step for diagnosing breast cancer in clinical practice. Nevertheless, masses can be partially obscured by high-intensity compacted glands especially in dense breasts, which imposes great challenges on radiologists and computer-aided detection (CAD) systems for the detection of mass from mammography. To better assist clinical diagnosis, mammogram mass detection is typically based on multiple views on both breasts. Specifically, as shown in Figure 1, a craniocaudal (CC) view (i.e., a top-down view of the breast) and a mediolateral oblique (MLO) view (i.e., a side view of the breast taken at a certain angle) are taken for both breasts. Comparing ipsilateral views (i.e. both CC and MLO views of the same breast) helps to analyze 3D structure of masses. Besides, since bilateral views (i.e., same view of both breasts) usually share a similar breast structure, asymmetric regions of bilateral views are more likely to be masses (Detailed definitions with respect to mammogram views will be described in Section 3). Therefore, the complementary effect of multi-view mammogram images is capable of providing valuable information regarding the breast anatomical prior structure, which is of great significance in digital mammography interpretation.

In terms of exploiting the relations of multi-view images, prior works can be roughly divided into two categories: ipsilateral view based and bilateral view based methods. For the ipsilateral view based modeling, an intuitive approach is to leverage the relation networks [3], [4] to model the ipsilateral inter-image non-local relations. For example, CVR-RCNN [5] cascades a relation module [3] to the second stage of Faster RCNN [6] to model the inter-proposal relations between CC and MLO views. However, compared to radiologists who can assist reasoning with domain knowledge, relation learning lacks clear constraints, i.e., the ipsilateral geometric and semantic relations are not explicitly taken into consideration. Thus, the learned relations may be incapable of precisely modeling the ipsilateral relations. In addition, it should be noted that such relation module relies, to a large extent, on the quality of region proposals at the first stage. When there exists the situation of severe gland occlusions, the performance will drop significantly. For the bilateral views [7], a recent work, i.e., CBN [8], proposes to fuse features of bilateral views with added tolerance of geometric distortions. However, like CVR-RCNN, CBN is based on RPN proposals [6], which also suffers from the proposal-missing problem.

During the exploration, we notice that radiologists can

* *indicates equal contribution.*
Corresponding author: Yizhou Yu.
Y. Liu, C. Chen, and S. Wang are with the AI Lab, Deepwise Healthcare, Beijing 100080, China.
F. Zhang is with the Center for Data Science, Peking University, Beijing 100871, China.
Y. Wang is with the Center on Frontiers of Computing Studies, Dept. of Computer Science & Technology, Advanced Institute of Information Technology, Peking University, Beijing 100871, China.
Y. Yu is with Department of Computer Science, The University of Hong Kong, Pokfulam, Hong Kong. E-mail: yizhouy@acm.org



(a) Examined view (b) Contralateral view (c) Auxiliary view (d) Zoom-in patches (e) Ideal projection model

Fig. 1. An illustration of the relation among mammography views. Standard mammography screening takes a CC view and an MLO view for each breast. Figure (a)-(c) represent the examined view (i.e., CC view of the right breast), the contralateral view (i.e., CC view of the left breast) and the auxiliary view (i.e., MLO view of the right breast) of a specific instance. The examined view represents the view where the detection is performed. Figure (a) and (b) form a bilateral pair, and are roughly symmetric since they have similar gland background, breast shape and breast size. Figure (a) and (c) form an ipsilateral pair, and provide complementary information to represent the 3D anatomical structure. Patch (1) and (3) in the corresponding images refer to a mass lesion instance, while Patch (2) locates similarly as Patch (1) in the contralateral image. Figure (d) offers zoom-in versions of these patches. Figure (e) stands for an ideal projection model of mammography. We can see that the CC view is a top-down view of the breast, while the MLO view is a side view taken at a certain angle along the pectoral muscle plane.

explicitly utilize the natural reasoning ability to identify masses by observing different mammographic views. The diagnosis flow adopted by the radiologists consists of the following three steps. (1) Determine the suspicious regions based on the examined view. (2) Search for compatible regions in the auxiliary view on the basis of appearances and locations, and contrast the corresponding regions based on the bilateral pair (i.e. examined view and contralateral view). (3) Make a diagnosis with respect to the suspicious regions according to two visual observations: (i) reasonable correspondences are found in the auxiliary view; and (ii) the regions in the bilateral view are symmetric. While the multi-view region-based reasoning procedure plays a key role for mammogram mass detection, most existing methods [8], [9], [10], [11], [12], [13], [14], [15], [16] focus on improving the detection accuracy in a single view. How to endow the existing object detection models with the capability of multi-view reasoning is vital for decision-making in clinical diagnosis but remains the boundary to explore.

Inspired by the aforementioned discussions, in this paper, we delve into the multi-view reasoning problem and propose an Anatomy-aware Graph convolutional Network (AGN), which is tailored for mammogram mass detection and endows detection models with multi-view reasoning ability. By jointly reasoning the correspondence relations among multiple mammogram views, AGN learns the intact multi-view information in an end-to-end manner during training. The input of AGN is the extracted features of multi-view images from the backbone network, and the output is the enhanced features of the examined view. As a general method, AGN can be easily plugged into any modern object detection frameworks [6], [17], [18] without modifying their original network architectures.

To be specific, the proposed AGN is comprised of the following three steps. **Firstly**, we introduce a novel Bipartite Graph convolutional Network (BGN) to model the *intrinsic geometric and semantic relations* of ipsilateral views. The construction of bipartite graph nodes aims at modeling the region-level correspondences between views. Each node represents a region with relatively consistent locations across breast instance. The bipartite graph edges are constructed to characterize the relation between nodes across views in two aspects: geometric constraints and appearance similarities. **Secondly**, considering that the visual asymmetry of bilateral views is widely adopted in clinical practice to assist the diagnosis of breast lesions [7], we propose an Inception Graph convolutional Network (IGN) to model the *structural similarities of bilateral views*. IGN targets on contrasting the bilateral mammogram views based on the assumption that asymmetric regions are more likely to be masses. Technically, it forms multi-branch graph connections between each node and its nearest neighbors, which strengthens the robustness of learned representations against inherent geometric distortions and forms an Inception-like structure [19]. **Finally**, based on the constructed bipartite and inception graphs, the multi-view information is propagated through nodes methodically after several layers of graph convolutions, which equips the features from the examined view with multi-view reasoning ability.

Compared to prior works that leverage weak or even no reasoning constraints, our AGN explicitly learns a customized multi-view reasoning model from both ipsilateral and bilateral views. In addition, the proposed BGN and IGN enhance the backbone features before the region proposal step, which helps to mitigate the proposal-missing problem. We evaluate the effectiveness of the proposed AGN on two mammogram mass detection benchmarks, i.e., a public dataset (DDSM [20]) and a multi-center in-house dataset. Our experiments reveal that the proposed algorithm signif-

icantly exceeds state-of-the-art performance on benchmark datasets. Moreover, visualization results demonstrate that the proposed AGN provides reasonable and interpretable visual cues for clinical diagnosis.

The contributions of our work are summarized as follows: Firstly, to the best of our knowledge, this is the first work that explicitly exploits the multi-view graphical correspondence for mammogram mass detection. Secondly, we propose a bipartite graph convolutional network, which is capable of performing reasoning about ipsilateral correspondences and modeling both geometric constraints and visual similarities between ipsilateral views. Lastly, we design an inception graph convolutional network, which models the structural similarities between bilateral views and enhances the robustness of learned representations relying on a *priori* that asymmetric regions are more likely to be masses.

This paper substantially extends our conference paper [21] from four major aspects. **(1)** Besides utilizing ipsilateral views [21], our AGN further considers the complementary effect of bilateral views to learn intact multi-view information of mammogram, which helps to make more comprehensive and precise clinical decisions. Specifically, we propose a novel inception graph convolutional network for modeling the structural similarities of bilateral views. **(2)** We enhance the mechanism of correspondence reasoning to fit the multi-view modeling scenario. **(3)** We conduct more experiments and ablation studies with respect to the enhanced network architecture, and include updated experimental results on a larger in-house multi-center dataset. **(4)** We provide more complete introduction and analysis for the proposed multi-view correspondence reasoning network, as well as more elaborated implementation details.

The rest of the paper is organized as follows: Section 2 gives a brief review of related works. Section 3 illustrates the preliminaries of mammogram views. Section 4 demonstrates the details of the proposed AGN. Experimental results and feature visualization are described in Section 5. Section 6 draws the conclusion.

2 Related Work

2.1 Mammogram Mass Detection

Existing works on mammogram mass detection [22] can be coarsely classified into two categories: traditional and deep-learning-based approaches. Traditional approaches rely on the handcrafted features to identify masses from mammogram images [23], [24]. The pipeline of these approaches usually includes two stages. The first stage, which aims at extracting region proposals to recall most masses, is to generate candidates. Based on the assumption that mass regions are brighter than background, region proposals are obtained using thresholding, clustering, bilateral image subtraction, etc [25], [26], [27], [28]. To further enlarge the intensity difference between mass regions and background, pre-processing methods are introduced, such as histogram equalization, exponent functions, etc. The second stage is to reduce the false positives. Numerous approaches [24], [29] resort to the handcrafted patterns to represent mass boundaries, textures or shapes. The traditional approaches suffer from the following limitations. Firstly, traditional approaches, which are based on handcrafted features, have weak representation ability and cannot be end-to-end trainable. Secondly, the generated candidates are very likely to contain many false positives, increasing the difficulty of optimizing the classifier in the second stage. Lastly, the second stage does not contain the localization step. Thus, the predicted locations of masses may largely deviate from their bounding boxes.

In the past decade, the renaissance in deep learning has greatly promoted the development of medical image computing [30], [31], [32], [33], [34], [35], [36], [37], [38], [39]. Mass detection has achieved remarkable success by virtue of deep convolutional neural networks [8], [9], [10]. A typical solution is to apply deep convolutional networks for reducing false positives [11], [12], [13]. However, these models cannot be end-to-end trainable, and thus result in inferior performance. To tackle this issue, researchers attempt to leverage the off-the-shelf modern object detectors, such as Faster R-CNN [6], FPN [40], Mask R-CNN [17], for the mammogram mass detection [14], [15], [16]. Despite their general efficacy, the complementary of multi-view mammogram images are not taken into considered. Ma *et al*. [5] propose to model the ipsilateral property and introduce a relation module [3] to the Faster RCNN [6], aiming to learn ipsilateral inter-proposal relations. However, the relation learning lacks clear constraints, i.e., the ipsilateral geometric and semantic relations are not explicitly considered. Thus, the learned relations may be incapable of precisely modeling the between-image correlations. In addition, such relation module heavily relies on the quality of region proposals at the first stage. When encountering the situation of severe gland occlusions, the detection performance will drop significantly. By explicitly leveraging the domain knowledge of specific image modalities, AGN has powerful multi-view reasoning ability, which significantly improves the localization ability of backbone features. MommiNet [41] is a concurrent work that proposes to simultaneously perform end-to-end bilateral and ipsilateral analysis of mammogram images. However, they do not consider the graphical correspondence among different mammographic views, which is important for the success of multi-view reasoning.

2.2 Visual Reasoning based on Graph Convolutional Network

Visual reasoning attempts to merge distinct information (interactions) among objects (scenes), and has been extensively explored in computer vision problems, e.g., image classification [42], [43], object detection [44], [45], semantic segmentation [46] and other visual understanding tasks [47], [48], [49], [50], [51]. The typical paradigm of visual reasoning is to incorporate the object relations or attributes into different vision tasks [3], [42], [52]. For example, Akata *et al*. [52] solve the attribute-based image classification problem by regarding it as a problem of reasoning in the attribute-embedded space.

Recently, Graph Convolutional Network (GCN) [53] has been introduced for visual reasoning tasks due to its representation power for non-Euclidean data and reasoning power from domain knowledge. Li *et al*. [46] propose a set of graph convolutional units to learning graph representations

from 2D visual data. However, the information propagation during the representation learning process will inevitably introduce noisy signals since all semantic correspondences are jointly considered. Moreover, the reasoning procedure is uncontrolled and implicit, which impairs the performance and the interpretability. Gao *et al*. [54] strengthen the expressive power of learned features for the visual tracking task by introducing a spatial-temporal GCN. However, the construction of graph nodes is based on the uniform grids, which are very sensitive to the variations of object scales, image shapes, geometric structures, etc. Considering the semantic dependencies among different objects, Xu *et al*. [55] propose to exploit the human commonsense knowledge to reason with a class-to-class prior. However, the constructed knowledge graph remains relatively fixed, and thus may fail to adapt to the more complex scenarios. In addition, it is incapable of extending to the single-category detection problems (e.g., mass detection). The reasoning procedure that radiologists read mammograms provides more explicit guidance, which motivates us to design a more customized algorithm with domain knowledge.

2.3 Multi-view Visual Recognition

Representing 3D object is one of the most fundamental problems in visual understanding [56], [57] and stereo vision [58], [59], [60], [61], [62]. Multi-view based approaches represent the 3D object as a collection of 2D views [63], [64]. Typically, they first conduct image-based classification on each individual view, and then aggregate the multi-view features for 3D representation. Feng *et al*. [57] propose a group-view CNN for hierarchical correlation modeling from multiple views. Yang *et al*. [65] propose a relation network which is capable of modeling region-to-region and view-to-view relationships from different viewpoints. For 3D medical images, such as computed tomography (CT) and magnetic resonance imaging (MRI), utilizing 2D image features projected from different views can lead to superior performance [66], [67]. For instance, Setio *et al*. [66] sample multiple views from the 3D CT image, and then cascade different classifiers trained on each individual view to boost the nodule classification performance.

Unlike multi-view based approaches in general computer vision [62] and 3D medical image analysis [66], [67], mammographic views have a differentiated imaging process, which requires us to design customized algorithms for analysis. Mammographic views are captured as the total absorption of all substances along the projection ray, which makes it impossible to decomposite the internal structures. Ipsilateral views are taken along different directions of the breasts, which provides richer information for representing the 3D structure of breast. Meanwhile, bilateral views have similar breast structure, and thus the asymmetric regions between views are more likely to be masses. In a nutshell, mammographic views owns more explicit correspondences, which motivates us to develop customized reasoning mechanisms. We note that the explicit correspondences are also be explored in stereo vision methods [59], [60], [61], [62], where they align key points via leveraging the explicit correspondences among calibrated cameras. In contrast to stereo vision, we can not obtain precisely matched correspondences between different views due to the existence of standard mammography screening protocols [68]. Thus, how utilize the fuzzy correspondences to enhance the expressive power of backbone features remains an open question.

3 Preliminary: Mammographic Views

In this section, we provide the details of the mammographic screening mechanism. As a special type of 2D radiography, digital mammographic images are captured as the total absorption of all substances along the projection ray. Therefore, using only a single view of mammographic images is insufficient to represent the breast internal structure. In standard mammographic screening, X-ray images are taken for both two breasts. For each breast, two mammographic views (i.e., CC view and MLO view) are taken by compressing the breast at a near orthogonal plane. Specifically, CC view is a top-down view while MLO view is a side view taken at a certain angle. Comparing ipsilateral views (i.e., both CC and MLO views of the same breast) helps to analyze the 3D structures of masses. Contrasting bilateral views (i.e., a specific view of both breasts) helps to extract suspicious mass lesions since bilateral views of breasts are approximately symmetric.

In this paper, a set of multi-view images are defined as input (illustrated in Figure 1), including an examined view (i.e., the view where the detection is performed), an auxiliary view (i.e., another view of the same breast) and a contralateral view (i.e., the view of opposite side of breast). We respectively utilize one of the mammogram images as the examined view, and the rest two views are defined as the auxiliary and contralateral views.

4 Methodology

4.1 Overview

The objective of AGN is to endow the mammogram mass detection framework with multi-view correspondence reasoning ability. By distilling multi-view information from the input multi-view mammogram images, AGN outputs the enhanced feature representations of the examined view for further detection. The overall architecture is shown in Figure 2, which consists of the following steps. (1) For the purpose of modeling the region-based reasoning procedure, the graph nodes are embedded into breasts, where each node represents the features of regions with relatively consistent locations in breasts. Then, bipartite graph convolutional network is introduced to model the geometric constraints and appearance similarities of nodes between ipsilateral views. (2) Inception graph convolutional network is designed to learn structural similarities of bilateral views with an added tolerance of geometric distortions. (3) Correspondence reasoning enhancement, which is based on the two pre-defined graph convolutional networks, is proposed to enhance the representation power of features. Based on the above steps, after information propagated through nodes, each node can not only be aware of the ipsilateral correspondences, but also learn the contrastive representations from bilateral views. It is noteworthy that the node representations are mapped to spatial visual domain reversely, which explicitly endows the spatial features with reasoning ability. In the

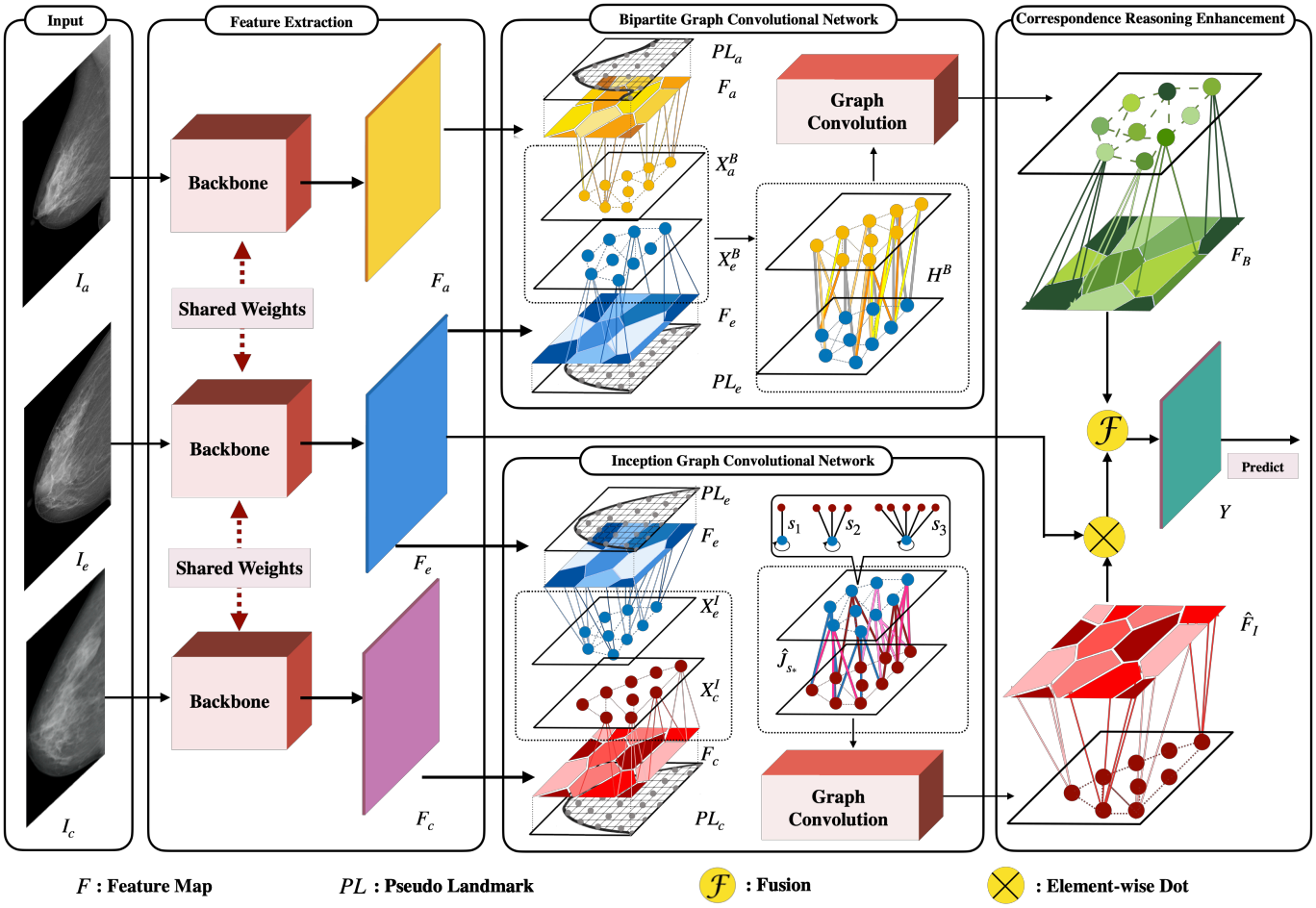


Fig. 2. The pipeline of the proposed AGN. AGN takes multi-view backbone features as inputs, and outputs enhanced features of the examined view for further prediction. First, bipartite graph convolutional network performs reasoning across ipsilateral views end outputs auxiliary representations of mass lesion 3D structure. Second, inception graph convolutional network contrasts bilateral views and produces attention maps on the suspicious asymmetric areas. Finally, correspondence reasoning enhancement based on the defined two graphs is conducted to enhance the backbone features of the examined view for further detection.

end, we fuse the enhanced features and the original backbone features for further proposals.

Specifically, we are given a set of 2D feature maps $F_e, F_a, F_c \in \mathbb{R}^{HW \times C}$ extracted from the examined view (simplified as e), the auxiliary view (simplified as a) and the contralateral view (simplified as c), and H, W and C represent the height, width and channel of the feature maps. $l_e, l_a, l_c \in \{CC, MLO\}$ are defined as view types. We guarantee that $l_e \neq l_a$ and $l_e = l_c$. As formulated in Equation 1, AGN learns a function f, parameterized by the bipartite graph $\mathcal{G}_B$ and the inception graph $\mathcal{G}_I$.

$$Y = f(F_e, F_a, F_c; \mathcal{G}_B, \mathcal{G}_I) \tag{1}$$

4.2 Graph Nodes

Graph nodes are introduced to denote the region-level correspondences in breasts with relatively consistent locations across different breast instances. Technically, we define graph nodes by considering two fundamental questions, i.e., "where to locate" and " what to represent".

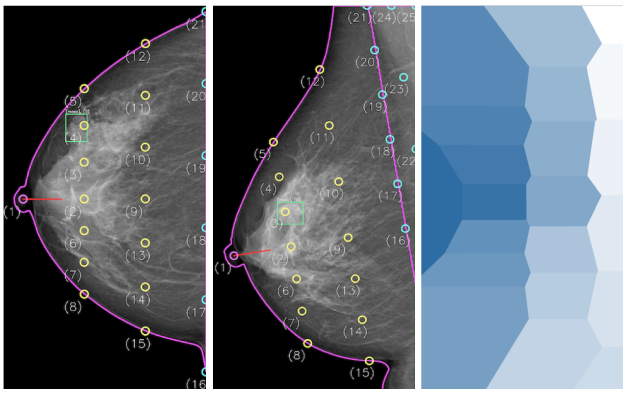
We introduce the concept of pseudo landmarks, which preserve relative consistent locations in breasts, to address the first question. For the second question, we note that the graph node mapping can produce node representations from spatial visual features. In the following parts, we provide the technical details.

4.2.1 Pseudo Landmarks

Landmarks represents points in a shape object, where correspondences between and within the populations of the object are preserved [69]. Unfortunately, there are no specialized landmarks for breasts, which inspires us to define pseudo landmarks based on prior knowledge.

The pseudo landmarks are expected to possess the following properties: I. Each pseudo landmark stands for a region with relatively consistent locations in breasts; II. Different pseudo landmarks should stand for distinct regions in breasts; III. Combining all pseudo landmarks are expected to cover the whole breast.

An intuitive approach is to regard uniform grids of the image as landmarks. However, property I. is not satisfied since the uniform grids are sensitive to the variations of image scale, geometric structures, etc. As demonstrated in Figure 1, the design principle of pseudo landmarks is based on a key observation: there exist clear geometric correspondences between CC and MLO views of standard mammog-



(a) CC View (b) MLO View (c) Mapping Cell

Fig. 3. Illustration of pseudo landmarks and bipartite graph node mapping. (a)-(b) draw pseudo landmarks and the matched bounding boxes on CC and MLO views respectively. (c) illustrates how bipartite node mapping works when $k = 1$. Each mapping cell denotes the representative region of the node in the CC view.

raphy screening. Ideally, a point in CC view approximately corresponds to a line in MLO view, which is parallel to the projected pectoral muscle plane.

As illustrated in Figure 3, in order to embed the pseudo landmarks, we first insert a set of equidistant parallel lines between the nipple and pectoral muscle line (projected by the pectoral muscle plane). Based on the intersection between parallel lines and breast contour, we uniformly insert points between two intersection points. After that, all the points are re-ordered based on the intersections and further defined as pseudo landmarks. Similarly, we define pseudo landmarks in pectoral muscle areas for MLO view. By doing so, we can obtain a set of pseudo landmarks for each view.

4.2.2 Graph Node Mapping

Graph node mapping targets on projecting spatial visual features $F \in \mathbb{R}^{HW \times C}$ to the node domain. Note that the features of each node are region-level features that represent a certain region in breasts.

The node mapping denotes the relation between a graph node and all pixels in the corresponding region. Formally, we design kNN (k Nearest Neighbor) forward mapping ϕ_k with its auxiliary matrix A for node feature representations. Each node is associated to an irregular region, satisfying the property that for any pixel in this region, the node is one of its k nearest nodes. ϕ_k performs region-level feature pooling within the regions corresponding to the graph nodes. The formulas are defined as follows:

$$\phi_k(F, \mathcal{V}) = (Q^f)^T F, \tag{2}$$

$$Q^f = A(\Lambda^f)^{-1}, \tag{3}$$

$$A_{ij} = \begin{cases} 1, & \text{if } j \text{ th node is kNN of } i \text{ th pixel.} \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

where $\mathcal{V}$ denotes the node set corresponding to spatial feature $F \in \mathbb{R}^{HW \times C}$, $A \in \mathbb{R}^{HW \times |\mathcal{V}|}$ is the auxiliary matrix that assigns spatial features to top-k nearest graph nodes, $\Lambda^f \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$ ($\Lambda^f_{jj} = \sum_{i=1}^{HW} A_{ij}$) is a diagonal matrix, and $Q^f \in \mathbb{R}^{HW \times |\mathcal{V}|}$ (a normalized form of A) denotes the forward mapping matrix.

Compared to fixed-grid assign methods [54], the node representations in our AGN are more robust to the variations of image scales, geometric structures, etc. The justification is that ϕ_k adaptively chooses representative region according to the relations among node locations. In addition, the proposed mapping mechanism has an explicit physical meaning, which has the merit of better visual interpretability. In particular, as show in figure 3 (c), the mapping degenerates to Voronoi grids [70] when $k = 1$.

4.3 Bipartite Graph Convolutional Network (BGN)

BGN learns to model ipsilateral relations among correspondences. BGN is characterized as $\mathcal{G}_B = (\mathcal{V}_{CC}, \mathcal{V}_{MLO}, \mathcal{E}_B)$, where $\mathcal{V}_{CC}$ and $\mathcal{V}_{MLO}$ represent the bipartite graph node sets constructed from CC view and MLO view respectively. $\mathcal{E}_B$ denotes the bipartite graph edge set. Each edge in $\mathcal{E}_B$ connects a node in $\mathcal{V}_{CC}$ to the corresponding one in $\mathcal{V}_{MLO}$, leading to a bipartite graph structure.

To obtain the feature representations of bipartite graph node, we resort to the kNN forward mapping described in Section 4.2.2. Formally, the bipartite graph node representations are defined as follows:

$$X^B_e = \phi_k(F_e, \mathcal{V}_{l_e}), \tag{5}$$

$$X^B_a = \phi_k(F_a, \mathcal{V}_{l_a}). \tag{6}$$

Here, for the sake of simplicity, we denote $X^{CC} \in \{X^B_e, X^B_a\}$ as node features for CC view and $X^{MLO} \in \{X^B_e, X^B_a\}$ for MLO view.

To obtain bipartite graph edge representations, we start by rethinking a fundamental question: what is the underlying relations between nodes? Given a mass which is located at one certain node in the examined view, it is clear that different nodes in the auxiliary view will have differentiated probabilities for representing the given mass. Motivated by this, we regularize the relation from two aspects, i.e., geometric constraints and appearance similarities. The two aspects characterize the inherent constraints resulted from mammogram screening mechanism and visual similarities between nodes, respectively.

Formally, bipartite graph edge is denoted as an adjacency matrix $H \in \mathbb{R}^{|\mathcal{V}_{CC}| \times |\mathcal{V}_{MLO}|}$, which consists of a geometric graph $H^g \in \mathbb{R}^{|\mathcal{V}_{CC}| \times |\mathcal{V}_{MLO}|}$ and a semantic graph $H^s \in \mathbb{R}^{|\mathcal{V}_{CC}| \times |\mathcal{V}_{MLO}|}$. The geometric graph is a global prior graph, which represents the geometric constraints across views. The semantic graph is an instance dependent graph, which characterizes the semantic similarities between nodes. The two graphs jointly regularize the ipsilateral information propagation. Eq. (7) demonstrates the relations of these two matrices,

$$H = H^g \circ H^s \tag{7}$$

where $\circ$ denotes the element-wise dot.

In the following subsections, we show the details of how to regularize the relations among correspondences.

4.3.1 Geometric Relation Learning

In this part, we explicitly model the geometric constraints. Even though the CC and MLO views own standard camera pose, it is difficult to define the precise geometric correspondence due to the tissue deformation and the lack of visual cues. To solve this issue, we resort to use masses as visual cues for modeling the geometric correspondence. Each edge in the geometric graph stands for the correlation of the linked nodes (i.e., the same mass instance from different views). In order to estimate the correlation, for each mass, if a node is the closest to the center of bounding box, this node will be selected to represent the mass. By doing so, we are capable of linking the nodes that stands for the same mass instance from different views (e.g., 4^{th} node in CC view and 3^{th} node in MLO view in Figure 3).

The construction of geometric graph H^g consists of the following two steps. (1) We construct a frequent statistics matrix ϵ for masses. If a node is the closest one to the center of the bounding box of a mass, this node is chosen to represent the mass. By doing so, we are capable of linking the nodes that stand for the same mass instance from different views. We traverse all labeled masses in the training set and obtain the frequent statistics matrix $\epsilon \in \mathbb{R}^{|\mathcal{V}_{CC}|\times|\mathcal{V}_{MLO}|}$. (2) To obtain H^g, we perform an augmented form of column-row normalization [55]:

$$H^g_{ij} = \frac{\epsilon_{ij}}{\sqrt{D_{i\cdot}.D_{\cdot j}}} \tag{8}$$

where $D_{i\cdot} = \sum_{k=1} \epsilon_{ik}$ and $D_{\cdot j} = \sum_{k=1} \epsilon_{kj}$.

4.3.2 Semantic Relation Learning

Whilst the geometric graph characterize the holistic geometric correlations, it still inevitably introduces noises during the reasoning procedure, and thus is hard to find exact correspondence pairs across views. In addition, we note that the appearance similarities between different views is a significant characteristic of mass lesions. Inspired by this, we introduce the semantic graph to learn the semantic relation between nodes, which is helpful for mitigating the negative influence of the noisy relations.

An intuitive approach to define the semantic similarities between nodes is to measure them by cosine similarity or inner product [71], [72]. However, the relations between nodes include the backgrounds, and their features may also be enhanced. To tackle this issue, we relax the weights, and allow the module to learn its own similarity as follows:

$$H^s_{ij} = \sigma([(X^{CC}_i)^T, (X^{MLO}_j)^T]w_s), \tag{9}$$

where $X^{CC}_i, X^{MLO}_j \in \mathbb{R}^C$ respectively denote the i^{th} and j^{th} node features of CC and MLO views, $w_s \in \mathbb{R}^{2C}$ stands for the fusion parameter, and σ denotes the sigmoid activation function.

4.4 Inception Graph Convolutional Network (IGN)

Based on the assumption that bilateral mammogram views share a similar breast structure and asymmetric regions are more likely to be suspicious regions, we propose the IGN to learn to contrast bilateral mammogram views. IGN links nodes with compatible locations from bilateral views and predicts attention values for regions in the examined view.

IGN is characterized as $\mathcal{G}_I = (\mathcal{V}_e \cup \mathcal{V}_c, \mathcal{E}_I)$, where $\mathcal{V}_e, \mathcal{V}_c$ indicate node sets constructed from the examined view and the contralateral view respectively. Since bilateral views have the same view type (i.e., $l_e = l_c$), we guarantee that $|\mathcal{V}_e| = |\mathcal{V}_c|$. For simplicity, we assume that $n = |\mathcal{V}_e| = |\mathcal{V}_c|$.

To obtain node feature representations for IGN, we adopt kNN forward mapping similarly. Formally, representations corresponding to the examined view and the contralateral view are defined as:

$$X^I_e = \phi_k(F_a, \mathcal{V}_e) \tag{10}$$

$$X^I_c = \phi_k(F_c, \mathcal{V}_c) \tag{11}$$

Then, the node representation of IGN can be defined as:

$$X^I = [(X^I_e)^T, (X^I_c)^T]^T \tag{12}$$

To obtain edge representation of IGN, we characterize the edge set $\mathcal{E}_I$ as an adjacency matrix $\hat{J} = \begin{pmatrix} M & J \\ J^T & M^T \end{pmatrix}$, which contains two components, i.e., $M \in \mathbb{R}^{n\times n}$ and $J \in \mathbb{R}^{n\times n}$. Specifically, M characterizes the relations of nodes within the same view, while J characterizes the relations of nodes across different views.

We set M to $\mathbf{0}$, indicating that there are no intra-connections within the view. However, determining the attention values of a certain view not only requires its contralateral information (i.e. J) but also the view information itself. Thus, we add self-loop for each node of the graph. Specifically, we set $M = I_n$.

As for the definition of J, it is intuitive to set $J = I_n$ which assumes that only nodes with the same location in bilateral views are linked. However, the bilateral views may not be aligned perfectly due to the inherent geometric distortions. To tolerate the distortions, we reformulate J as J_s which links each node to its top-s nearest neighbors (NN) in the contralateral view. J_s provides larger visual context which helps to increase the tolerance for the distortions. Note that each distinct value of s can induce a distinct set of cross-view edges and their corresponding cross-view adjacency matrix is denoted as J_s.

The set of cross-view edges corresponding to J_s can be regarded as a single branch of graph linkages. Instead of using a single branch, we adopt multiple branches of graph linkages, each corresponding to a distinct cross-view adjacency matrices. This kind of GCN can provide stronger representation abilities and form an Inception-like structure [19]. Specifically, supposing that IGN has two different branches $s_1, s_2 \in \mathbb{N}^*$, the induced corresponding augmented adjacency matrices are denoted as $\hat{J}_{s1}, \hat{J}_{s2}$. When performing graph convolutions, both $\hat{J}_{s_1}$ and $\hat{J}_{s_2}$ affect information propagation. Details of convolution operation will be described in Section 4.5.2.

4.5 Correspondence Reasoning Enhancement

Correspondence reasoning enhancement, which is based on the defined bipartite graph $\mathcal{G}_B$ and the inception graph $\mathcal{G}_I$, is developed to fully explore multi-view reasoning procedure for enhancing the customized features. It consists of the following steps. (1) Augment bipartite graph convolution, making it adapt to the modern graph convolutional manner;

(2) Design inception graph convolution with multi-branch connections among nodes; (3) Map node representations to spatial domain reversely after several layers of graph convolutions; and (4) Fuse the original backbone features with the graph representations learned from BGN and IGN to enhance the expressive power of final representations. We will describe the details in the following subsections.

4.5.1 Bipartite Graph Convolution

To adapt the modern GCN [54], [73], we provide the augmented form of the bipartite graph:

$$X^B = [(X^{CC})^T, (X^{MLO})^T]^T, \tag{13}$$

$$H^B = \begin{pmatrix} \mathbf{0} & H \\ H^T & \mathbf{0} \end{pmatrix}, \tag{14}$$

where $X^B \in \mathbb{R}^{|\mathcal{V}_{CC} \cup \mathcal{V}_{MLO}| \times C}$ denotes the augmented form of the bipartite graph nodes, and $H^B \in \mathbb{R}^{|\mathcal{V}_{CC} \cup \mathcal{V}_{MLO}| \times |\mathcal{V}_{CC} \cup \mathcal{V}_{MLO}|}$ is the augmented form of the adjacency matrix.

We follow the common practice [54] to define graph convolution. An iteration of graph convolution layer is defined in Equation 15, where $W^B \in \mathbb{R}^{C \times C}$ and σ indicate the convolution parameters and sigmoid activation function. To this end, we are able to stack multiple layers to form the the graph convolutional network.

$$Z^B = \sigma(H^B X^B W^B) \tag{15}$$

4.5.2 Inception Graph Convolution

To achieve inception graph convolution with multi-branch linkage among nodes, we generalize standard graph convolution operations. We give the formulation of an iteration of the inception graph convolutional operation in Equation 16. Feature transformations of multiple branches are conducted independently, and then the transformed multi-branch features are aggregated. For simplicity, the equation contains only two branches. It is intuitive to reformulate it to adapt to multi-branch settings.

$$Z^I = \sigma\left(\left(\hat{J}_{s_1} \hat{J}_{s_2}\right) \begin{pmatrix} X^I & \mathbf{0} \\ \mathbf{0} & X^I \end{pmatrix} \begin{pmatrix} W_1^I \\ W_2^I \end{pmatrix}\right), \tag{16}$$

where $W_1^I, W_2^I \in \mathbb{R}^{C \times C}$ indicate parameters of the layer.

4.5.3 kNN Reverse Mapping

In order to enhance the spatial features, we introduce a kNN reverse mapping function ψ_k to map the graph node features to the spatial domain. Following the design principle of the kNN forward mapping (cf. Section 4.2.2), we keep the same number (k) of nearest neighbors. ψ_k is defined as :

$$\psi_k(Z, \mathcal{V}_e) = Q^r [Z]_e, \tag{17}$$

$$Q^r = (\Lambda^r)^{-1} A, \tag{18}$$

where Z denotes the node presentations after graph convolutions, $\mathcal{V}_e$ stands for the node set from the examined view, $A \in \mathbb{R}^{HW \times |\mathcal{V}_e|}$, which is correspond to $\mathcal{V}_e$, is defined by following Equation 4, $[\cdot]_e$ denotes an indexing operator which chooses nodes in the examined view from all nodes, $\Lambda^r \in \mathbb{R}^{HW \times HW}$ represents a diagonal matrix, $\Lambda^r_{ii} = \sum_{j=1}^{|\mathcal{V}_e|} A_{ij}$, and $Q^r \in \mathbb{R}^{HW \times |\mathcal{V}_e|}$ denotes the reverse mapping matrix which is the normalized form of A.

TABLE 1
Performance on DDSM dataset(%).

Method	R@t
Campanini *et at*. [74]	80@1.1
Eltonsy *et at*. [75]	92@5.4, 88@2.4, 81@0.6
Sampat *et at*. [76]	88@2.7, 85@1.5, 80@1.0
Faster RCNN [5]	85@2.1, 75@1.8, 73@1.2
CVR-RCNN [5]	92@4.4, 88@1.9, 85@1.2
AG-RCNN	**96@4.4, 92@1.9, 90@1.2**

TABLE 2
Performance on DDSM dataset(%).

Method	R@0.5	R@1.0	R@2.0	R@3.0	R@4.0
Faster RCNN, FPN	75.3	81.5	87.3	89.8	91.4
Faster RCNN, FPN, DCN	75.7	82.5	88.4	90.1	91.4
Mask RCNN, FPN	76.0	82.5	88.7	90.8	91.4
Mask RCNN, FPN, DCN	76.7	83.9	89.4	91.4	91.8
BG-RCNN [21]	79.5	86.6	91.8	92.5	94.5
AG-RCNN	**82.0**	**89.0**	**92.1**	**93.8**	**95.5**

4.5.4 Feature fusion

Feature fusion procedure includes the following steps. We first map the features from the auxiliary view to spatial domain using ψ_k. After that, F_B and F_e are aligned to the same coordinate space, which helps to fuse them more effectively:

$$F_B = \psi_k(Z^B, \mathcal{V}_e) \tag{19}$$

Then, we predict attention values for regions in the examined view induced by inception graph convolutions.

$$F_I = \psi_k(Z^I, \mathcal{V}_e), \tag{20}$$

$$\hat{F}_I = \sigma(F_I w_I), \tag{21}$$

where $w_I \in \mathbb{R}^C$ represents the parameter, and $\hat{F}_I$ refers to the spatial attention map. Finally, we enhance the features based on the processed feature triple, which has been aligned in the same spatial coordinate space:

$$Y = [\hat{F}_I \cdot F_e, F_B] W_f^T, \tag{22}$$

where $\cdot$ indicates spatial-wise dot which broadcasts along channel axis, $W_f \in \mathbb{R}^{C \times 2C}$ represents the fusion parameter.

5 Experiments

5.1 Implementation Details

In experiments, the mammogram images are segmented by OTSU [77], and we use the foreground regions as the input. In order to keep the same spatial resolution among different view of images, each view of input image is resized to same size of the examined image. We leverage the Hough transform to detect the pectoral muscle line and nipple in three steps for pseudo landmark embedding. First, points potentially lying on the pectoral muscle line are extracted with the Canny edge detector. Then, these extracted points are mapped into the parameter space. Finally, noisy points in the parameter space are removed according to the prior location of the pectoral muscle line, and the optimal point in the parameter space is identified for the pectoral muscle

line. The point on the breast contour with the largest distance to the detected pectoral muscle line is further located as the nipple. We apply several specific data augmentation methods, such as random flipping, random cropping, and multi-scaling, to prevent over-fitting during the training stage.

The proposed AGN is integrated into Mask RCNN [17] with ResNet-50 [78] architecture, and we term the full mammogram mass detection framework as AG-RCNN hereafter. The parameters of ResNet-50 are fine-tuned from the model pre-trained on ImageNet. The loss function follows the same definition as Mask RCNN [17], containing three terms: classification loss, regression loss, and segmentation loss. The FPN anchors span 5 scales and 3 aspect ratios [40]. To adapt to modern FPN [40] network structure, AGN enhances each level of feature pyramid with shared parameters, since node representations are invariant to feature map scales.

Our experiments are implemented by the PyTorch deep learning framework [32]. We utlize stochastic gradient descent (SGD) for the training with a learning rate 0.02, weight decay 10^{-4}, momentum 0.9 and nesterov set True. The training process takes 30 epochs in all. Regarding the BGN, we keep the same number of nearest neighbors k for both ϕ_k and ψ_k for bipartite node mapping and reverse mapping. As for IGN, the number of nearest neighbors k for both ϕ_k and ψ_k is set to 1 for keeping higher spatial resolutions.

5.2 Datasets

We perform experiments on both a public dataset DDSM [20] and an in-house dataset. Note that we do not choose other public datasets (such as INBreast [79] , MIAS [80]). The justification is that the sample size of these datasets is insufficient to train a detection model.

DDSM dataset. DDSM dataset includes 2620 mammography cases, and most cases contain two views of images for both breasts. Following previous practices [5], [74], [75], [76], the DDSM dataset is divided to 1897 cases for training, 211 cases for validation and 512 cases for testing.

In-house dataset. The in-house dataset includes 10,000 cases, which are collected from four different vendors: IMS s.r.l., Siemens, Hologic, and GE Healthcare. Each case contains a CC view and an MLO view for each breast, and thus there are 40,000 images in all. The annotations, namely the mask of each mass lesion, are labeled by 3 radiologists with strong expertise. If there are disagreements among them, we will adopt the majority opinion of radiologists. This dataset is randomly split into training, validation and testing sets in a ratio of 8:1:1.

5.3 Baselines

Faster RCNN, FPN. Faster RCNN [6] with Feature Pyramid Network (FPN) [40] is a strong baseline in object detection task. FPN enhances the modeling ability of detecting multi-scale objects. It assigns objects to different level of feature maps according to object scales. We use ResNet-50 [78] as the backbone network.

TABLE 3
Performance on in-house dataset(%).

Method	R@0.5	R@1.0	R@2.0	R@3.0	R@4.0
Faster RCNN, FPN	82.3	85.4	90.5	92.5	93.7
Faster RCNN, FPN, DCN	83.1	88.0	91.0	92.5	93.9
Mask RCNN, FPN	83.1	88.0	91.4	93.4	94.2
Mask RCNN, FPN, DCN	84.0	88.3	91.7	93.2	94.5
BG-RCNN [21]	85.7	89.4	92.1	93.7	95.0
AG-RCNN	**87.6**	**90.6**	**93.4**	**94.7**	**95.2**

Faster RCNN, FPN, DCN. Deformable Convolution Network (DCN) [81] is proposed to boost the transformation modeling capability of convolutional networks. DCN is introduced into the baselines to further enhance the detection performance.

Mask RCNN, FPN, DCN. Mask RCNN [17] is a state-of-the-art approach for both object detection and instance segmentation tasks. To leverage mask annotations for precise localization, we exploit Mask RCNN framework for boosting the performance.

Mask RCNN, FPN, DCN. DCN is further integrated into the Mask RCNN baselines to improve the performance.

CVR-RCNN. CVR-RCNN [5] models the ipsilateral relations of mammograms by adding a relation module [3] into the second stage of detection process.

5.4 Comparison with State-of-the-art Methods

We evaluate the performance by recall (R) at t ($t \in \{0.5, 1.0, 2.0, 3.0, 4.0\}$) false positive per image (FPI), which is simplified as $R@t$. A mass region is recalled when its IOU (Intersection Over Union) value is larger than 0.2.

Table 1 and Table 2 display the experimental results on DDSM dataset. Baseline results in Table 1 are cited from their original papers [5], [21], [74], [75], [76]. In Table 2, we re-implemented the baseline methods in our experiments. We do not make a comparison with [8] since their split of dataset is different. We keep the same FPI and compare with a strong baseline approach, CVR-RCNN [5]. We can see that AG-RCNN significantly outperforms all compared methods. The results on in-house dataset are reported in Table 3. Although in-house dataset has larger amount of data and image modalities, our approach consistently outperforms all comparison methods, which verifies the effectiveness and robustness of AG-RCNN on the challenging scenario.

MommiNet [41] is an existing work on multi-view mammogram mass detection but with different experimental settings. Following the same practice of MommiNet, we randomly divide all cases on the DDSM dataset into the training, validation, and test sets by approximately 8:1:1, resulting in 8,256, 1,020 and 1,036 images in the respective sets. The proposed AG-RCNN outperform MommiNet by +1.5% (@0.5), +1.8% (@1.0), and +2.3% (@2.0).

To explore how the proposed model benefits from the correspondence reasoning mechanism, we qualitatively analyze some cases in Figure 4. As shown in the 2nd and 3rd

TABLE 4
Effectiveness of pseudo landmarks on DDSM and In-house datasets (%).

Method	DDSM					In-house				
	R@0.5	R@1.0	R@2.0	R@3.0	R@4.0	R@0.5	R@1.0	R@2.0	R@3.0	R@4.0
Uniform Grids	77.7	85.6	91.5	93.2	94.5	85.6	89.4	81.9	93.0	94.1
Proposed	**82.0**	**89.0**	**92.1**	**93.8**	**95.5**	**87.6**	**90.6**	**93.4**	**94.7**	**95.2**

TABLE 5
Effectiveness of node number on DDSM and In-house datasets (%).

Method	DDSM					In-house				
	R@0.5	R@1.0	R@2.0	R@3.0	R@4.0	R@0.5	R@1.0	R@2.0	R@3.0	R@4.0
$PL(1,1)$	77.0	84.2	91.4	92.5	93.2	84.5	89.6	92.1	94.2	94.7
$PL(9,13)$	80.5	87.1	91.6	93.5	94.5	86.4	90.1	92.6	94.5	95.0
$PL(21,25)$	81.1	87.0	91.8	93.2	94.5	87.2	90.0	92.9	94.3	95.0
$PL(42,46)$	81.4	87.3	92.0	93.8	95.2	87.1	90.4	93.2	**94.9**	95.2
$PL(66,71)$	**82.0**	**89.0**	**92.1**	**93.8**	**95.5**	**87.6**	**90.6**	**93.4**	94.7	**95.2**

TABLE 6
Effectiveness of graph node mapping on DDSM and In-house datasets (%).

Method	DDSM					In-house				
	R@0.5	R@1.0	R@2.0	R@3.0	R@4.0	R@0.5	R@1.0	R@2.0	R@3.0	R@4.0
$PL(21,25)$, Crop	78.8	86.3	91.1	92.5	94.2	85.4	88.4	91.4	93.2	94.2
$PL(21,25)$, $k=1$	80.8	86.8	91.4	92.5	94.2	86.2	90.2	92.5	94.5	95.1
$PL(21,25)$, $k=2$	81.1	87.0	91.8	93.2	94.5	87.2	90.0	92.9	94.3	95.0
$PL(21,25)$, $k=3$	80.1	86.3	92.1	93.0	94.2	86.6	90.3	92.8	94.2	94.7
$PL(66,71)$, Crop	78.4	87.0	92.1	92.5	93.5	86.1	89.0	92.8	93.5	95.0
$PL(66,71)$, $k=1$	80.1	87.0	92.5	93.5	94.9	86.7	90.3	92.9	94.6	95.0
$PL(66,71)$, $k=2$	80.5	87.7	92.5	93.8	95.2	87.2	90.2	93.2	94.2	95.0
$PL(66,71)$, $k=3$	**82.0**	**89.0**	**92.1**	**93.8**	**95.5**	**87.6**	**90.6**	**93.4**	**94.7**	**95.2**

TABLE 7
Ablation of components in bipartite graph convolutional network on DDSM and in-house datasets (%).

H^g	H^s	In-house					DDSM				
		R@0.5	R@1.0	R@2.0	R@3.0	R@4.0	R@0.5	R@1.0	R@2.0	R@3.0	R@4.0
		76.7	83.9	89.4	91.4	91.8	84.0	88.3	91.7	93.2	94.5
√		77.7	86.3	89.4	91.8	93.5	84.7	88.9	92.1	93.4	94.0
	√	78.4	83.9	91.1	92.1	93.8	85.2	89.2	92.0	93.4	94.2
√	√	**79.5**	**86.6**	**91.8**	**92.5**	**94.5**	**86.2**	**89.5**	**92.5**	**93.6**	**94.6**

TABLE 8
Ablation of components in inception graph convolutional network on DDSM and In-house datasets (%).

Method	DDSM					In-house				
	R@0.5	R@1.0	R@2.0	R@3.0	R@4.0	R@0.5	R@1.0	R@2.0	R@3.0	R@4.0
IGN(1)	78.0	84.5	90.4	92.1	93.8	84.5	89.5	92.4	94.1	94.9
IGN(3)	78.1	86.0	91.0	92.5	93.5	85.0	89.0	92.3	94.2	94.4
IGN(5)	78.4	85.3	91.1	93.2	94.2	85.3	89.1	92.3	94.1	94.7
IGN(1, 3)	78.8	85.8	91.1	93.2	94.2	85.2	89.3	92.4	94.2	94.9
IGN(1, 3, 5)	**79.1**	**86.3**	**91.4**	**93.2**	**94.5**	**85.6**	**89.5**	**92.8**	**94.4**	**95.0**

TABLE 9
Ablation of modules on DDSM and in-house datasets (%).

BGN	IGN	DDSM					In-house				
		R@0.5	R@1.0	R@2.0	R@3.0	R@4.0	R@0.5	R@1.0	R@2.0	R@3.0	R@4.0
		84.0	88.3	91.7	93.2	94.5	76.7	83.9	89.4	91.4	91.8
√		86.2	89.5	92.5	93.6	94.6	79.5	86.6	91.8	92.5	94.5
	√	85.6	89.5	92.8	94.4	95.0	79.1	86.3	91.4	93.2	94.5
√	√	**87.6**	**90.6**	**93.4**	**94.7**	**95.2**	**82.0**	**89.0**	**92.1**	**93.8**	**95.5**

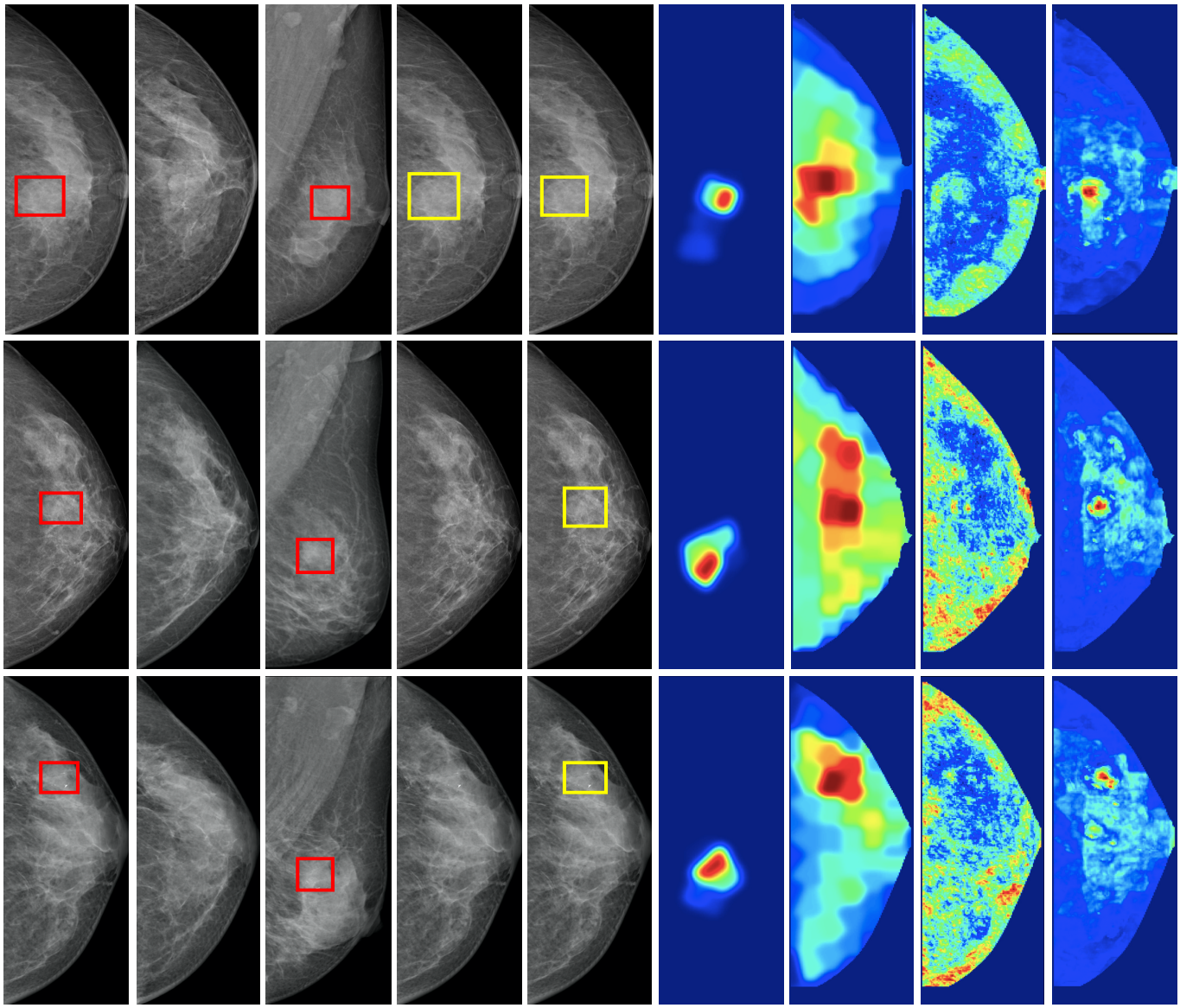


Fig. 4. Detection results of AG-RCNN. Each row shows a representative case. Column (a)-(c) refer to the examined view, the flipped contralateral view and the auxiliary view with annotations. Column (d)-(e) indicate detection results by Mask-RCNN and AG-RCNN. Column (f) visualizes the attention area on the auxiliary view. Column (g) shows the attention regions of bilateral views. Column (h)-(i) visualize the response maps before and after correspondence reasoning enhancement.

rows, detecting mass lesions with only a single view is quite confusing since mass lesions are obscured by compacted glands in breasts. Leveraging visual cues from different mammographic views can provide clear and reasonable evidences for mass detection in the examined view, and thus make the detection process more efficient and interpretable. By doing so, the proposed method can significantly improve the recall. Besides, as shown in 1st row, the localization of the bounding box becomes more precise, since more detailed information about the anatomical structure of mass is taken into consideration.

5.5 Ablation Study

5.5.1 Ablation of Pseudo Landmarks

We compare the proposed pseudo landmarks with uniform grids, which embeds nodes uniformly into mammogram images without considering the visual prior knowledge. The number of nodes is identical between uniform grids and pseudo landmarks. The results are shown in Table 4, which clearly demonstrate the superiority of our pseudo landmarks. We further investigate the effect of the number of nodes on the performance. "$PL(x, y)$" in Table 5 denotes the setting that there are x nodes in CC view and y nodes in MLO view. The setting "$PL(1, 1)$" is approximatively equivalent to two-branch Faster RCNN. The results are reported in Table 5, we choose "$PL(66, 71)$" as our final results.

5.5.2 Ablation of Graph Node Mapping.

To explore the effectiveness of bipartite node mapping, we first compare with a simple baseline, which directly crops a fixed region for graph node representation. Results in Table 6 verify the superiority of the proposed graph node mapping. The justifications can be summarized as follows. First, the graph node mapping endows node representations with the property that the representative region of each node is invariant to the variations of image scales and breast shapes, which enhances the robustness of the node representation. Second, the graph node mapping mechanism is easy-to-implement without any additional parameters, which simplifies the training process.

We further analyze the relationship between k and the results. We keep k fixed for both ϕ_k and ψ_k. As shown in Table 6, when the dense nodes are embedded, the model performs better with a larger k. The reason is that the dense nodes have smaller representative regions, and a larger k can extract richer context features for each node.

5.5.3 Ablation of Bipartite Graph Convolutional Network.

We conduct the ablation study of bipartite graph convolutional network based on ipsilateral views. To analyze the influence of each component in the bipartite graph convolutional network, we isolate each component (i.e., H^g and H^s) of bipartite graph edges. We traverse all combinations of using H^g and H^s. It degenerates to a plain Mask RCNN when neither H^s nor H^g are used. The justification is that no information propagates across views and the prediction is only based on the examined view. When either H^s or H^g is used, we set H to H^s or H^g respectively, which propagates information across correspondences satisfying semantic constraints or geometric constraints only. The results are shown Table 7, which reveal two critical observations. First, compared with the single view based method, utilizing either H^s or H^g can provide considerable improvements. Second, by combining semantic and geometric relations, we can achieve the best performance.

5.5.4 Ablation of Inception Graph Convolutional Network

We conduct the ablation study of inception graph convolutional network based on bilateral views. For simplicity, "IGN(s_1, s_2, s_3)" denote IGN with the settings of three different branches, i.e., s_1, s_2, s_3. Table 8 reports the experimental results, which demonstrates the following conclusions. First, equipping the model with the tolerance of geometric distortions will enhance the performance. Second, adopting multi-branch information propagation among top-s nearest nodes achieves the best performance.

5.5.5 Ablation of Modularities

We evaluate all combinations of BGN and IGN. The model degenerates to a plain Mask RCNN if neither BGN nor IGN are used, since the detection is only based on the examined view without multi-view information propagation. When using BGN only, the enhanced feature Y is reformulated as:

$$Y = [F_e, F_B]W_f^T. \tag{23}$$

While using IGN, the enhanced feature Y with parameter $W_f \in \mathbb{R}^{C\times C}$ can be reformulated as:

$$Y = (\hat{F}_I \cdot F_e)W_f^T. \tag{24}$$

The results are summarized in Table 9, which demonstrate that the performance gain benefits from both BGN and IGN.

5.6 Visualization

Our visualization experiments mainly answer three questions: (1) Where does the bipartite graph focus on auxiliary view? (2) Where does inception graph convolutional network focus on bilateral views? (3) How does the correspondence reasoning mechanism enhance the feature representations?

Firstly, we develop a specialized method for correspondence visualization to answer the first question. The major objective is to seek the representative regions of correlated nodes in the auxiliary view when given a query mass in the examined view. We define a one-hot representative vector $x \in \mathbb{R}^{|\mathcal{V}_{CC}\cup\mathcal{V}_{MLO}|}$ to denote the locations of the query masses in the examined area. The index of the node, which is nearest to the center of the analyzed mass in the examined view, is set to 1. We visualize the feature via Equation 25, where $o \in \mathbb{R}^{HW}$ stands for the response vector, and $[\cdot]_e$ represents indexing operator which selects nodes in the examined view from bipartite graph node set. We reshape and normalize the response vector o to the output image.

$$o = Q^r[H^B x]_e \tag{25}$$

As can be seen in Figure 4 (f), we found that the bipartite graph focuses on the matched mass area in the auxiliary view, which is helpful for learning complementary feature representations. In addition, the proposed model has a clear physical meaning and provides visual cues of matched masses. Therefore, it is capable of assisting radiologists in clinical mammography interpretation.

Secondly, we visualize the attention regions learned by the inception graph convolutional network. Since there exists a natural spatial attention map $\hat{F}_I$, we simply normalize it and obtain the visualization map. As illustrated in Figure 4 (g), attention regions mostly appear at asymmetric areas in bilateral views, which provides positive evidence of the regions to be mass lesions.

Lastly, to investigate how correspondence reasoning mechanism enhances the feature representations, we compare the response map before and after feature enhancement. To be specific, we respectively conduct channel-wise max pooling on F_e and Y. The results are shown in Figure 4. We can observe that feature response map activates more prominently on the mass region after enhancement. By doing so, the corresponding reasoning enhancement method helps to promote the detection performance and make a sufficient and comprehensive clinical decision.

6 Conclusion

In this paper, we delve into the multi-view correspondence reasoning problem and introduce a anatomy-aware graph convolutional network to endow the mammogram mass detection models with customized reasoning ability. By

jointly reasoning and distilling information from multiple mammography views, our model substantially enhances the expressive power of learned representations in the examined view during the detection process. The proposed model includes a bipartite graph convolutional network and an inception graph convolutional network. The former one is capable of performing reasoning about ipsilateral correspondences and modeling both geometric constraints and visual similarities across ipsilateral views, and the latter one can model the structural similarities between bilateral views. To this end, correspondence reasoning enhancement propagates information through both graphs, which makes the spatial visual features aware of the multi-view correspondences. Extensive experiments on both public and in-house datasets reveal that the proposed model significantly exceeds the state-of-the-art performance. In addition, visualization results show that AGN provides reasonable and interpretable visual cues for the clinical diagnosis.

ACKNOWLEDGMENT

This work was supported in part by Zhejiang Province Key Research & Development Program (No. 2020C03073), MOST-2018AAA0102004, NSFC-61625201, 61527804, DFG TRR169 / NSFC Major International Collaboration Project "Crossmodal Learning".

REFERENCES

[1] R. Siegel, J. Ma, Z. Zou, and A. Jemal, "Cancer statistics, 2014," CA: a cancer journal for clinicians, vol. 64, no. 1, pp. 9–29, 2014.

[2] E. A. Sickles, "Breast cancer screening outcomes in women ages 40-49: clinical experience with service screening using modern mammography," JNCI Monographs, vol. 1997, no. 22, pp. 99–104, 1997.

[3] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3588–3597.

[4] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks."

[5] J. Ma, S. Liang, X. Li, H. Li, B. H. Menze, R. Zhang, and W.-S. Zheng, "Cross-view relation networks for mammogram mass detection," arXiv preprint arXiv:1907.00528, 2019.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in Advances in neural information processing systems, 2015, pp. 91–99.

[7] H. Chen, Y. Wang, K. Zheng, W. Li, C.-T. Cheng, A. P. Harrison, J. Xiao, G. D. Hager, L. Lu, C.-H. Liao et al., "Anatomy-aware siamese network: Exploiting semantic asymmetry for accurate pelvic fracture detection in x-ray images," arXiv preprint arXiv:2007.01464, 2020.

[8] Y. Liu, Z. Zhou, S. Zhang, L. Luo, Q. Zhang, F. Zhang, X. Li, Y. Wang, and Y. Yu, "From unilateral to bilateral learning: Detecting mammogram masses with contrasted bilateral network," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2019, pp. 477–485.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[10] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzebski, T. Févry, J. Katsnelson, E. Kim et al., "Deep neural networks improve radiologists' performance in breast cancer screening," 2019.

[11] T. Kooi, G. Litjens, B. Van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, and N. Karssemeijer, "Large scale deep learning for computer aided detection of mammographic lesions," Medical image analysis, vol. 35, pp. 303–312, 2017.

[12] N. Dhungel, G. Carneiro, and A. P. Bradley, "A deep learning approach for the analysis of masses in mammograms with minimal user intervention," Medical Image Analysis, vol. 37, pp. 114–128, 2017.

[13] J. O. B. Diniz, P. H. B. Diniz, T. L. A. Valente, A. C. Silva, A. C. de Paiva, and M. Gattass, "Detection of mass regions in mammograms by bilateral analysis adapted to breast density using similarity indexes and convolutional neural networks," Computer methods and programs in biomedicine, vol. 156, pp. 191–207, 2018.

[14] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, "Detecting and classifying lesions in mammograms with deep learning," Scientific reports, vol. 8, no. 1, p. 4165, 2018.

[15] H. Jung, B. Kim, I. Lee, M. Yoo, J. Lee, S. Ham, O. Woo, and J. Kang, "Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network," PloS one, vol. 13, no. 9, p. e0203355, 2018.

[16] Z. Cao, Z. Yang, X. Zhuo, R.-S. Lin, S. Wu, L. Huang, M. Han, Y. Zhang, and J. Ma, "Deeplima: Deep learning based lesion identification in mammograms," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 0–0.

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[18] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "Reppoints: Point set representation for object detection," arXiv preprint arXiv:1904.11490, 2019.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[20] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer, "The digital database for screening mammography," in Proceedings of the 5th international workshop on digital mammography. Medical Physics Publishing, 2000, pp. 212–218.

[21] Y. Liu, F. Zhang, Q. Zhang, S. Wang, Y. Wang, and Y. Yu, "Cross-view correspondence reasoning based on bipartite graph convolutional network for mammogram mass detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[22] H.-D. Cheng, X. Shi, R. Min, L. Hu, X. Cai, and H. Du, "Approaches for automated detection and classification of masses in mammograms," Pattern recognition, vol. 39, no. 4, pp. 646–668, 2006.

[23] N. R. Mudigonda, R. M. Rangayyan, and J. L. Desautels, "Detection of breast masses in mammograms by density slicing and texture flow-field analysis," IEEE Transactions on Medical Imaging, vol. 20, no. 12, pp. 1215–1227, 2001.

[24] S.-C. Tai, Z.-S. Chen, and W.-T. Tsai, "An automatic mass detection system in mammograms based on complex texture features," IEEE journal of biomedical and health informatics, vol. 18, no. 2, pp. 618–627, 2014.

[25] D. Brzakovic, X. M. Luo, and P. Brzakovic, "An approach to automated detection of tumors in mammograms," IEEE Transactions on Medical Imaging, vol. 9, no. 3, pp. 233–241, 1990.

[26] A. J. Méndez, P. G. Tahoces, M. J. Lado, M. Souto, and J. J. Vidal, "Computer-aided diagnosis: Automatic detection of malignant masses in digitized mammograms," Medical Physics, vol. 25, no. 6, pp. 957–964, 1998.

[27] H. Li, Y. Wang, K. R. Liu, S.-C. Lo, and M. T. Freedman, "Computerized radiographic mass detection. i. lesion site selection by morphological enhancement and contextual segmentation," IEEE Transactions on Medical Imaging, vol. 20, no. 4, pp. 289–301, 2001.

[28] L. Zhen and A. K. Chan, "An artificial intelligent algorithm for tumor detection in screening mammogram," IEEE transactions on medical imaging, vol. 20, no. 7, pp. 559–567, 2001.

[29] J. Wei, B. Sahiner, L. M. Hadjiiski, H.-P. Chan, N. Petrick, M. A. Helvie, M. A. Roubidoux, J. Ge, and C. Zhou, "Computer-aided detection of breast masses on full field digital mammograms," Medical physics, vol. 32, no. 9, pp. 2827–2838, 2005.

[30] Z. Xu, Y. Huo, J. Park, B. Landman, A. Milkowski, S. Grbic, and S. Zhou, "Less is more: Simultaneous view classification and landmark detection for abdominal ultrasound images," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2018, pp. 711–719.

[31] Z. Guo, L. Zhang, L. Lu, M. Bagheri, R. M. Summers, M. Sonka, and J. Yao, "Deep logismos: Deep learning graph-based 3d seg-

mentation of pancreatic tumors on ct scans," in 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, 2018, pp. 1230–1233.

[32] N. Dhungel, G. Carneiro, and A. P. Bradley, "Automated mass detection in mammograms using cascaded deep learning and random forests," in 2015 international conference on digital image computing: techniques and applications (DICTA). IEEE, 2015, pp. 1–8.

[33] F. Zhang, L. Luo, X. Sun, Z. Zhou, X. Li, Y. Yu, and Y. Wang, "Cascaded generative and discriminative learning for microcalcification detection in breast mammograms," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12 578–12 586.

[34] S. Wang, M. Zhou, Z. Liu, Z. Liu, D. Gu, Y. Zang, D. Dong, O. Gevaert, and J. Tian, "Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation," Medical image analysis, vol. 40, pp. 172–183, 2017.

[35] F. Yellin, B. D. Haeffele, S. Roth, and R. Vidal, "Multi-cell detection and classification using a generative convolutional model," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8953–8961.

[36] J. Liu, G. Zhao, Y. Fei, M. Zhang, Y. Wang, and Y. Yu, "Align, attend and locate: Chest x-ray diagnosis via contrast induced attention network with limited supervision," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 10 632–10 641.

[37] B. Wu, X. Sun, L. Hu, and Y. Wang, "Learning with unsure data for medical image diagnosis," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 10 590–10 599.

[38] Z. Li, C. Wang, M. Han, Y. Xue, W. Wei, L.-J. Li, and L. Fei-Fei, "Thoracic disease identification and localization with limited supervision," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8290–8299.

[39] W. Shen, M. Zhou, F. Yang, D. Yu, D. Dong, C. Yang, Y. Zang, and J. Tian, "Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification," Pattern Recognition, vol. 61, pp. 663–673, 2017.

[40] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.

[41] Z. Yang, Z. Cao, Y. Zhang, M. Han, J. Xiao, L. Huang, S. Wu, J. Ma, and P. Chang, "Momminet: Mammographic multi-view mass identification networks," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2020, pp. 200–210.

[42] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Word spotting and recognition with embedded attributes," IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 12, pp. 2552–2566, 2014.

[43] K. Marino, R. Salakhutdinov, and A. Gupta, "The more you know: Using knowledge graphs for image classification," arXiv preprint arXiv:1612.04844, 2016.

[44] X. Chen, L.-J. Li, L. Fei-Fei, and A. Gupta, "Iterative visual reasoning beyond convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7239–7248.

[45] C. Jiang, H. Xu, X. Liang, and L. Lin, "Hybrid knowledge routed modules for large-scale object detection," in Advances in Neural Information Processing Systems, 2018, pp. 1552–1563.

[46] Y. Li and A. Gupta, "Beyond grids: Learning graph representations for visual recognition," in Advances in Neural Information Processing Systems, 2018, pp. 9225–9235.

[47] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009, pp. 951–958.

[48] I. Misra, A. Gupta, and M. Hebert, "From red wine to red tomato: Composition with context," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1792–1801.

[49] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in Advances in neural information processing systems, 2013, pp. 2121–2129.

[50] J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille, "Learning like a child: Fast novel visual concept learning from sentence descriptions of images," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2533–2541.

[51] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[52] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 819–826.

[53] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, and M. Sun, "Graph neural networks: A review of methods and applications," arXiv preprint arXiv:1812.08434, 2018.

[54] J. Gao, T. Zhang, and C. Xu, "Graph convolutional tracking," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4649–4659.

[55] H. Xu, C. Jiang, X. Liang, L. Lin, and Z. Li, "Reasoning-rcnn: Unifying adaptive global reasoning into large-scale object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6419–6428.

[56] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas, "Volumetric and multi-view cnns for object classification on 3d data," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5648–5656.

[57] Y. Feng, Z. Zhang, X. Zhao, R. Ji, and Y. Gao, "Gvcnn: Group-view convolutional neural networks for 3d shape recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 264–272.

[58] A. Kar, C. Häne, and J. Malik, "Learning a multi-view stereo machine," in Advances in neural information processing systems, 2017, pp. 365–376.

[59] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3260–3269.

[60] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[61] X. Chen, K. Kundu, Y. Zhu, H. Ma, S. Fidler, and R. Urtasun, "3d object proposals using stereo imagery for accurate object class detection," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 5, pp. 1259–1272, 2017.

[62] H. Rebecq, G. Gallego, E. Mueggler, and D. Scaramuzza, "Emvs: Event-based multi-view stereo—3d reconstruction with an event camera in real-time," International Journal of Computer Vision, vol. 126, no. 12, pp. 1394–1414, 2018.

[63] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 945–953.

[64] E. Johns, S. Leutenegger, and A. J. Davison, "Pairwise decomposition of image sequences for active multi-view recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3813–3822.

[65] Z. Yang and L. Wang, "Learning relationships for multi-view 3d object recognition," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7505–7514.

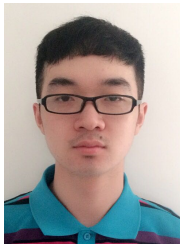
[66] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. Van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sánchez, and B. van Ginneken, "Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks," IEEE transactions on medical imaging, vol. 35, no. 5, pp. 1160–1169, 2016.

[67] F. Ciompi, B. de Hoop, S. J. van Riel, K. Chung, E. T. Scholten, M. Oudkerk, P. A. de Jong, M. Prokop, and B. van Ginneken, "Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2d views and a convolutional neural network out-of-the-box," Medical image analysis, vol. 26, no. 1, pp. 195–202, 2015.

[68] M. P. Sampat, M. K. Markey, A. C. Bovik et al., "Computer-aided detection and diagnosis in mammography," Handbook of image and video processing, vol. 2, no. 1, pp. 1195–1217, 2005.

[69] I. L. Dryden and K. V. Mardia, Statistical Shape Analysis, with Applications in R. Second Edition. Chichester: John Wiley and Sons, 2016.

[70] F. Aurenhammer and R. Klein, "Voronoi diagrams," Handbook of computational geometry, vol. 5, no. 10, pp. 201–290, 2000.
[71] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images." in Iccv, vol. 98, no. 1, 1998, p. 2.
[72] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 2. IEEE, 2005, pp. 60–65.
[73] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.
[74] R. Campanini, D. Dongiovanni, E. Iampieri, N. Lanconelli, M. Masotti, G. Palermo, A. Riccardi, and M. Roffilli, "A novel featureless approach to mass detection in digital mammograms based on support vector machines," Physics in Medicine & Biology, vol. 49, no. 6, p. 961, 2004.
[75] N. H. Eltonsy, G. D. Tourassi, and A. S. Elmaghraby, "A concentric morphology model for the detection of masses in mammography," IEEE transactions on medical imaging, vol. 26, no. 6, pp. 880–889, 2007.
[76] M. P. Sampat, A. C. Bovik, G. J. Whitman, and M. K. Markey, "A model-based framework for the detection of spiculated masses on mammography a," Medical physics, vol. 35, no. 5, pp. 2110–2123, 2008.
[77] N. Otsu, "A threshold selection method from gray-level histograms," IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62–66, Jan 1979.
[78] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[79] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "Inbreast: toward a full-field digital mammographic database," Academic radiology, vol. 19, no. 2, pp. 236–248, 2012.
[80] P. SUCKLING J, "The mammographic image analysis society digital mammogram database," Digital Mammo, pp. 375–386, 1994.
[81] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 764–773.



Yuhang Liu received the Bachelor degree in Computer Science and Technology from Xidian University in 2016, and the Master degree in Intelligence Science and Technology from Peking University in 2019. He is a machine learning researcher at Deepwise AI Lab. His research interests include biomedical image analysis, deep learning and computer vision.

Fandong Zhang is currently a post-doc in Academy for Advanced Interdisciplinary Studies at Peking University. Previously, he received the Ph.D. degree in EECS from Peking University. His research interests include medical image analysis, computer vision and biometrics. Recently, he is working on automatic analysis of breast medical imaging, including mammography, MRI and Ultrasound.

Chaoqi Chen received the B.S. and M.E. Degrees from Xiamen University, China, in 2017 and 2020, respectively. He is currently a Research Intern at Deepwise AI Lab, Beijing, China. His research spans computer vision and machine learning, with special interests in deep learning, transfer learning, and their visual applications.

Siwen Wang received the BS and MS degrees in information and communication engineering from Dalian University of Technology, Dalian, China, in 2017 and 2019, respectively. She is currently with the Deepwise AI Lab, Beijing, China. Her current research interests include computer vision, pattern recognition, and medical image analysis.

Yizhou Wang is a BoYa Professor of Computer Science Department and the vice director of the Center on Frontiers of Computing Studies at Peking University. He received his Bachelors degree in Electrical Engineering from Tsinghua University in 1996, and his Ph.D. in Computer Science from University of California at Los Angeles (UCLA) in 2005. He joined Xerox Palo Alto Research Center (Xerox PARC) as a research staff from 2005 to 2007. He was granted the National Natural Science Fund (NSFC) for Distinguished Young Scholars. Dr. Wang's research interests include computational vision, statistical modeling and learning, medical image analysis, and digital visual arts.

Yizhou Yu (M'10, SM'12, F'19) received the PhD degree from University of California at Berkeley in 2000. He is a professor at the University of Hong Kong, and also the chief scientist at Deepwise Healthcare. He was a faculty member at University of Illinois at Urbana-Champaign for twelve years. He is a recipient of 2002 US National Science Foundation CAREER Award and ACCV 2018 Best Application Paper Award. Prof Yu has served on the editorial board of IET Computer Vision, The Visual Computer, and IEEE Transactions on Visualization and Computer Graphics. He has also served on the program committee of many leading international conferences, including CVPR, ICCV, and SIGGRAPH. His current research interests include computer vision, deep learning, biomedical data analysis, computational visual media and geometric computing.